National Collaboratories and Data Management

Richard P. Mount

Stanford Linear Accelerator Center

The success of the National Collaboratories program within SciDAC has made it a strong candidate model for any program focused on the advancement of computer science in support of collaborative science.

In the first half of this year, DOE-MICS has supported a series of Data-Management Workshops studying data-management issues that affect Office of Science Programs. The workshops revealed that Office of Science programs are broadly challenged by the volume and complexity of their data, and that these programs offer a unique environment in which the science of data-intensive science can be advanced.

The current National Collaboratories program is already closely connected to the world of collaborative data management through projects such as the Particle Physics Data Grid. In the future, science can only gain from a close relationship between the National Collaboratories program and a program funding cross-disciplinary work on data management issues.

What would such a program look like? The key similarity with the National Collaboratories program would be a focus on a partnership between applications and computer science producing almost immediate benefits within the framework of a coherent long-term vision. The Data-Management Workshops showed huge areas of overlap between the needs of application sciences. These overlaps were most striking when the scientific activities were viewed as “simulation-driven”, “experiment or observation-driven”, or “information intensive”. Clearly, any “data-management collaboratory” program should take this into account in the organization of its application-focused activities.

A more technological view of the data-management problem also found many common themes. The workshops divided the technical and computer-science issues into 6 areas:

1. Workflow, dataflow, data transformation
2. Metadata, data description and logical organization
3. Efficient access and queries, data integration
4. Distributed data management, data movement, networks
5. Storage and caching
6. Integrated Data Analysis and Visualization Environments

Some work is already being funded in all these areas. Work may be in the context of SciDAC including the National Collaboratories program, or may be the efforts of disciplines like high-energy and nuclear physics to stay afloat in a tidal wave of data. Some effort is even funded through the MICS base program. In general, existing work

and funding fall well short of a optimal program exploiting commonalities between the sciences.

Which needs for data-management technology are vital to science, which are less important but still essential components of cost-effective science? The Workshops developed a psychological approach to try to answer these questions honestly. Each group of application scientists was maneuvered into imagining that they were paying for the desirable R&D out of their own program funds. Each was offered just four FTEs to work on data-management issues and was asked to assign these to the most pressing tasks. Combining effort with other application sciences was encouraged. The result was startling – a clear picture of priorities emerged, whereas when asked simply to rate topics of low, medium or high importance the result was an unfundable number of "highs".

The priorities established by a workshop at one point in time cannot possibly have total long-term validity. Today's bottlenecks will be removed and new ones will appear. New application-science directions will generate unexpected data-management needs. Therefore an approach to prioritization for data-management R&D must be found that continues to give the right answers as programs evolve.

The principle demonstrated by the psychological approach adopted at the workshops is already alive and well in the National Collaboratories program and is certainly the right approach to funding data-management R&D. Application science must resoundingly endorse the value of proposed R&D by being prepared to contribute to its funding, and the application's interest in and control over what happens to its resources can be assured by making the R&D a collaborative effort between applications and computer science. The presence of application sciences as stakeholder is the best way to ensure that any program evolves according to their scientific priorities.

The range of sciences becoming information-enabled and information-challenged is growing rapidly. The timing is perfect for an organized collaborative approach to the exploration of the science of data-intensive science. Such an approach will enable, and sometimes liberate, scientific exploration while providing advances in computer science that will have national and international impact on commerce and government.